ICTNET at Blog Track TREC 2010

Xueke Xu[1,2], Yue Liu[1], Hongbo Xu[1], Xiaoming Yu[1], Zeying Peng[1,2], Xueqi Cheng[1], Lihao Xiao[1,2], Shuaishuai Nie[3]

1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 100190

2. Graduate School of Chinese Academy of Sciences, Beijing, China, 100190

3. Department of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, 430074

ABSTRACT

This paper describes our participation in blog track of TREC2010. We submit runs for both two tasks, this paper mainly describe approaches to the two tasks.

1. Introduction

The Blog Track 2009 introduced two pilot tasks, i.e. faceted blog distillation and top stories identification, and each task has two separate sub-tasks. The Blog Track 2010 refines the two tasks of 2009 in many aspects. For example, one major change is that a two-stage submission strategy is adopted which facilitates separately investigating the performance and robustness of deployed approaches for the second sub-task. In this year ICTNET group participates in blog track and submits runs for both two tasks.

For both tasks, data preprocessing, which mainly focuses on post content extraction, plays an important role, and we adopt a link tables removing algorithm [5] to detect valuable content blocks from post pages and discard noisy blocks. The blog track use a collection called Blogs08 which is one order of magnitude bigger than Blogs06 and amounts to over 2TB of data, making indexing and retrieveling more challenging. We use "FirteX" platform[1], which is developed by our lab, for indexing and retrieveling preprocessed posts.

For blog distillation sub-task, inspired by the idea of "ensemble ranking", we combine various rankings to improve the robustness of our system. These rankings may differ from each other in underlying representation models, pseudo-relevance feedback approaches or resources used for pseudo-relevance feedback.

For faceted blog distillation sub-task, a language model is learnt for each facet inclination using Google blog search service and annotated data by Know-center. Based on the learnt language model, a generation model is introduced to combine topic-relevance and facet inclination degree in a probabilistic framework. With this model, baseline rankings without considering any facet feature are improved for faceted sub-task, and are further combined to get the final faceted run.

For story ranking sub-task, we use training data crawled from Reuters website to learn a classifier to categorize news stories into 5 categories, and then we measure the importance of each news story by accumulating the BM25 scores of posts published on the query day, treating headline and content of the story as query respectively.

For news blog post ranking sub-ask, there are two runs without special consideration of diversity requirement, we adopt a similar "ensemble ranking" strategy with blog distillation task for these two runs. For another run considering diversity, we explicitly extract and model aspects of each news story based on k-means clustering technology, and posts with formerly well covered aspects are more penalized in the ranking procedure.

[1] http://www.firtex.org/

2. Faceted Blog Distillation Task

2.1 Baseline Sub-Task

Candidate feeds selection

For each query, we first select candidate feeds based on the assumption that a relevant feed should contain at least one relevant post. To this end, for each topic, we produce a list of top N ad-hoc relevant posts based on our "FirteX" platform. The relevance scores of posts are computed according to a variant of BM25 model which takes into account proximity information of query words [2]. We use this list to identify candidate feeds. The parameter N is set to 2500 according to the training on the 2009 topics. By feeds selection, we prune feeds not deserving to be ranked, and thus remarkably improve the efficiency of our system.

Based on this candidate feed set, we can produce various rankings for baseline sub-task. We may appropriately select and combine these rankings to improve the robustness of our system. Basically, these rankings may fall into two kinds according to their underlying feed representation models: Local Representation Model and Global Representation model.

Local Representation Model

In this model, each feed is considered as a collection of its constituent posts. How to accumulate the individual post evidence to infer the feed's overall relevance is a key issue. To this end, we adopt small document model which exploits the relationship between the post and the feed [1]. In this model the topic relevance of feed F is given by the likelihood of F given Q as follows.

$$P_{SD}(F \mid Q) = P(F)\sum_{P \in F} P(Q \mid P)P(P \mid F)$$

Here, $P(F)$ is feed prior, which is $\log(N_F)$ in our system, flavoring large feeds, N_F is the size of feed F, $P(Q \mid P)$ is query likelihood of post P measuring topic relevance, and $P(P \mid F)$ is post centrality.

Model components are estimated in various ways, and correspondingly, we have following rankings:

1. $P(Q \mid P)$ is given by the retrieval score by the variant of BM25, $P(P \mid F)$ is uniform for each post.
2. Similar with 1, but $P(P \mid F)$ is proportional to exponential value of negative KL divergence between respective language mode. Both two models are estimated using Maximum Likelihood Estimation.
3. Similar with 1，but $P(Q \mid P)$ is given by the classic BM25. Corresponding query is the expansion words obtained with a pseudo-relevance feedback approach based on Wikipedia Resource, and expansion words are weighted with Bol model [3].

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
ICTNET at Blog Track TREC 2010	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Chinese Academy of Sciences,Institute of ComputingTechnology,Beijing, China, 100190,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Nineteenth Text REtrieval Conference (TREC 2010) held in Gaithersburg, Maryland on 16-19 November 2010. The conference was co-sponsored by the National Institute of Standards and Technology (NIST), the Defense Advanced Research Projects Agency (DARPA), and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
This paper describes our participation in blog track of TREC2010. We submit runs for both two tasks, this paper mainly describe approaches to the two tasks.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**3**	

It's notable that we only exploit top N ad-hoc relevant posts (i.e. $P(Q|P)$ is set to 0 if post P is not in the list). The reason may be that since most posts are irrelevant or weakly relevant even they contains some query terms, they may play an overwhelming part in a feed, weaken the information delivered by relevant posts and make model biased to noisy information.

Global Representation Model

This model considers a blog as a virtual document which is concatenation of all its constituent posts. This model may avoid words sparsity issue and reflect the recurring interest of the blog, but ignore any distinction between individual posts within the feed.

We use a language model approach to rank. Specifically, both the feed and the query are represented as a language model (i.e. multinomial distribution over words), respectively, and relevance score is negative KL divergence between the two language models.

Feed language model can be inferred using Maximum Likelihood Estimation with Dirichlet smoothing, while query language model can be inferred by pseudo-relevance feedback based approaches. Given a query, we use different resources (internal or external) for obtaining pseudo-relevance document collection, and based on the obtained collection we use different word weighting approaches, measuring how informative the word is in pseudo-relevance collection against the whole collection, to infer the probability of a word. Correspondingly, we have following different rankings.

4. Blogs08 Resource, Bol word weighting approach
5. Blogs08 Resource, Divergence Minimization word weighting approach [6]
6. Wikipedia Resource, Bol word weighting approach
7. Google blog search service Resource, Bol word weighting approach

We also incorporate temporal evidence into the final ranking score. We adapt the idea of entropy to measure whether a feed has a recurring interest in given Q.

$$Recurring\text{-}Degree(F,Q)=\frac{-\sum_{t=1}^{m}\frac{r_t(F,Q)}{\sum_{i=1}^{M}r_i(F,Q)}\log(\frac{r_t(F,Q)}{\sum_{i=1}^{M}r_i(F,Q)})}{log(M)}$$

Here, M is the number of days throughout the timespan, $r_t(F,Q)$ is the sum of relevance scores of constituent posts published on day t, $\log(M)$ is used for normalization. To obtain final ranking score, the relevance score will be multiplied by the Recurring-Degree score for all above rankings.

Rankings Combination

Given a list of rankings: $rks=\{rk_1,rk_2,...,rk_n\}$, let $pos(rk_m,F)$ be the position of feed F in the ranking rk_m ,the combination ranking of rks be $comb(rks)$, then the final ranking score of feed F in $comb(rks)$ can be computed as:

$$\exp(\frac{-\sum_{rk\in rks}pos(rk,F)}{|rks|}) .$$

Note that our combination manner can be hierarchical (i.e. a combination ranking may be further combined with other rankings). Two combination strategies are obtained by training in the 2009 topics, and correspondingly, we have two different baseline runs.

2.2 Faceted Blog Distillation Sub-Task

For faceted blog distillation sub-task, we introduce a generative model which combines topic-relevance and facet inclination degree in a probabilistic framework. In this model, faceted ranking score of feed F for facet inclination V is given by the likelihood of F given Q and V_Q , where V_Q is topic-specific facet inclination language model:

$$P(F|Q,V_Q)=\sum_w P(F|Q,w)P(w|V_Q)$$

We can easily derive like that:

$$P(F|Q,V_Q)=P(F)P(Q|F)\sum_w P(w|V_Q)P(w|F,Q)$$

Obviously, there are two parts in this model. $P(F)P(Q|F)$ reflects topic relevance of the feed, and can be estimated using any formerly mentioned approaches to baseline sub-task . $\sum_w P(w|V_Q)P(w|F,Q)$ gives facet inclination degree estimation, where $P(w|V_Q)$ is probability of word w in V_Q , $P(w|F,Q)$ is probability of w given F and Q, which is estimated depending on both query Q and feed F. Specifically, we estimate it with Maximum Likelihood Estimation, but only considering topic-relevant part of feed F (i.e. only top 2500 relevant posts of query Q) to highlight words closely related to the topic.

Via this model, we combine two factors of topic relevance and facet inclination degree to infer ranking score of each feed in a probabilistic framework with theoretical justification.

For each facet inclination V, we assign a weight to each word, which reflect its relatedness with specific facet inclination in general, using annotated data by Know-Center[4]. Then, given a query, we can learn a topic-specific language model V_Q by following steps:

- First, we submit the original query to Google blog search service, and fetch top 100 topical relevant web pages.
- Second, we use top weighted words as query to compute a score for each page using BM25 model. Top 30 pages are used as pseudo-facet-inclination-relevance pages.
- Finally, we use Bol word weighting approach, measuring how informative the word is in the pseudo-facet-inclination-relevance page set against the whole Blogs08 collection, to infer the probability of a word in V_Q .

$P(F)P(Q|F)$ can be estimated using formerly mentioned approaches to baseline sub-task. Correspondingly, we can obtain different faceted rankings by replacing $P(F)P(Q|F)$ with ranking scores in the baseline rankings, respectively. We also use the V_Q LM for ranking feeds by the negative KL divergence, and get one more faceted ranking.

Similar with baseline sub-task, we obtain faceted runs by selecting and combining these faceted rankings.

3. Top Stories Identification Task

3.1 Story Ranking Sub-Task

We first use training data crawled from Reuters website to learn a classifier to categorize news stories into 5 categories. Then, based on the observation that important news stories should be those concerning wide-ranging influential events and thus mentioned by bloggers extensively, we measure the importance of a news story by summing up the BM25 relevance scores of posts on given day, treating its headline or content as the query respectively. Specifically, for a given query day *d*, the importance of a news story *N* can be measured by following formula:

$$Score_{content}(N,d) = \sum_{post \in d} rel_{BM25}(N_{content}, post)$$

$$Score_{headline}(N,d) = \sum_{post \in d} rel_{BM25}(N_{headline}, post) \quad ,$$

where

$$rel_{BM25}(N_{content}, post) = \sum_{w \in N_{content}} P(w|N_{content}) \cdot BM25(w, post)$$

$P(w|N_{content})$ is the probability of word w in news content language model inferred by Bol model. $Score_{headline}(N,d)$ is computed likewise.

Finally, for each category, we rank news stories belonging to that category according to their importance. Note that for a category which has no enough stories, we add to the ranking list the news stores for which the category is second-likely according to the classifier with corresponding importance scores discounted.

3.2 News Blog Post Ranking Sub-Task

There are two runs without special consideration of diversity criterion, we adopted a similar "ensemble ranking" strategy with blog distillation task for these two runs. For each news story, we first retrieve top 50000 blog posts relevant to the news story using headline as query with classic BM25 model. Among these posts, we only consider the blog posts with timestamp >= query timestamp -2 and <= query timestamp + 9. Since posts concerning the event may be issued around the event day with a burst characteristic, and posts deviating from this day are highly probably irrelevant. Then we estimate news story language model with following different approaches, respectively.

1. Use news content and Bol word weighting approach to infer the probability of word *w* in news story language model
2. Use top 15 posts in original retrieval results and Bol model to infer the probability of word w in news story language model.
3. Similar with 2, but a post sharing common feed with formerly picked posts will be discounted for its contribution to the language model
4. Similar with 2, but top 15 posts are obtained according to the negative KL divergence between news story content language model by 1 and post language model. Here post language model is estimated using MLE.
5. Similar with 4, but a post sharing common feed with formerly picked posts will be discounted for its contribution to the language model.

As the task require, there should be three rankings which are centered at a different period of time respectively. For each period, we choose candidate posts within the required period, and then we compute ranking score for each post with the negative KL divergence between the news story language model and the post language model. Post language model is estimated using MLE. Finally, we have different rankings corresponding to different news story language models. Note that the original retrieval ranking is also considered for combination.

Similar with blog distillation task, we obtain the two runs by selecting and combining these rankings.

For the run considering diversity, we explicitly extract and model aspects of each news story based on k-means clustering technology, and posts will be penalized for sharing formerly covered aspects in the ranking procedure. Specifically, we partition top 150 posts into 5 disjoint clusters using k-means clustering technology. We assume each cluster represent an aspect of the news story. Then, we adopt a greedy strategy to iteratively pick support posts. First, ranking score of each candidate post is initiated with its combination score. Then, at each iteration step, top scored unpicked post is picked, and the ranking score of each unpicked post is penalized according to aspect distribution similarity between the post and the formerly picked posts.

4. ACKNOWLEDGMENTS

We thank Thomson-Reuters for kindly providing TRC2 newswire corpus, and thank Know-Center for providing annotated blog data. This work was funded by National Natural Science Foundation of China under grant number 60903139, 60933005. 973 Program of China 2007CB311103.

5. REFERENCES

[1]Elsas, J. L., Arguello, J., Callan, J., and Carbonell, J. G. 2008. Retrieval and feedback models for blog feed search. In *Proceedings of SIGIR 2008*.

[2]Guan, F., Yu, X., Peng, Z., Xu, H., Liu, Y., Song, L., and Cheng, X. ICTNET at Web Track 2009 Ad-hoc task. In *Proceedings of TREC-2009*.

[3]He, B., Macdonald, C., He, J., and Ounis, I. 2008. An effective statistical approach to blog post opinion retrieval. In *Proceeding of CIKM '08*.

[4]Lex, E., Granitzer, M., Muhr, M., and Juffinger, A. Stylometric features for emotion level classification in news related blogs. In *Proceedings of the 9th RIAO Conference (RIAO 2010), 2010*.

[5]Song, L., Cheng, X., Guo, Y., Liu, Y., and Ding, G. 2009. ContentEx: a framework for automatic content extraction programs. In *Proceedings of the 2009 IEEE international Conference on intelligence and Security informatics*.

[6]Zhai, C. and Lafferty, J. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM '01*.